ÖAW AUSTRIAN ACADEMY OF SCIENCES

Managing the Lifecycle of Research Data –
A Crash Course

4th AI Winter School of the ÖAW
22.01.2025

Content

Introduction

- What are data and research data?
- Data in AI and ML
- Data and metadata
- Data Lifecycle

Data Management

- Why?
- Data Management Plan
- FAIR data

Core Tasks Areas of Data Management

Key Takeaways

Who are we?

Data Management, Digital Preservation & Archiving Team aka PANDA

ARCHE

arche.acdh.oeaw.ac.at

Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)

Massimiliano

Seta

Martina

Introduction

Better Data is Better Than Better Models

firstfinger · Follow
3 min read · Nov 11, 2023

4

Machine Learning

Data-Centric AI: AI Models Are Only as Good as Their Data Pipeline

To produce higher performing AI systems, researchers need to focus on the data.

Jan 25, 2022 | Katharine Miller

THE AI JOURNAL

TOPICS ADVERTISE AWARDS LEARN COMPANY WRITE FOR US Log In

Home / Topics / AI / Analytics / The limits of AI: A model is only as good as its data

Analytics

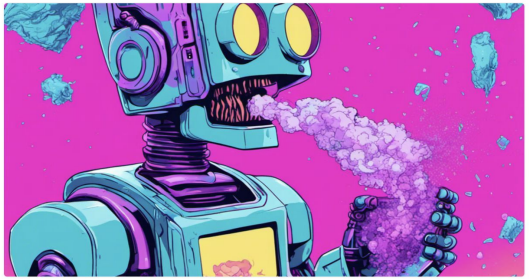
The limits of AI: A model is only as good as its data

Hannah Algar · October 9, 2024 · 6 minutes read

NOVEMBER 1, 2023

AI is only as good as the data: Q&A with Satish Jayanthi of Coalesce

AI systems obey the golden rule: garbage in, garbage out, Want good results, feed it good data.



Credit: StackPlusOne

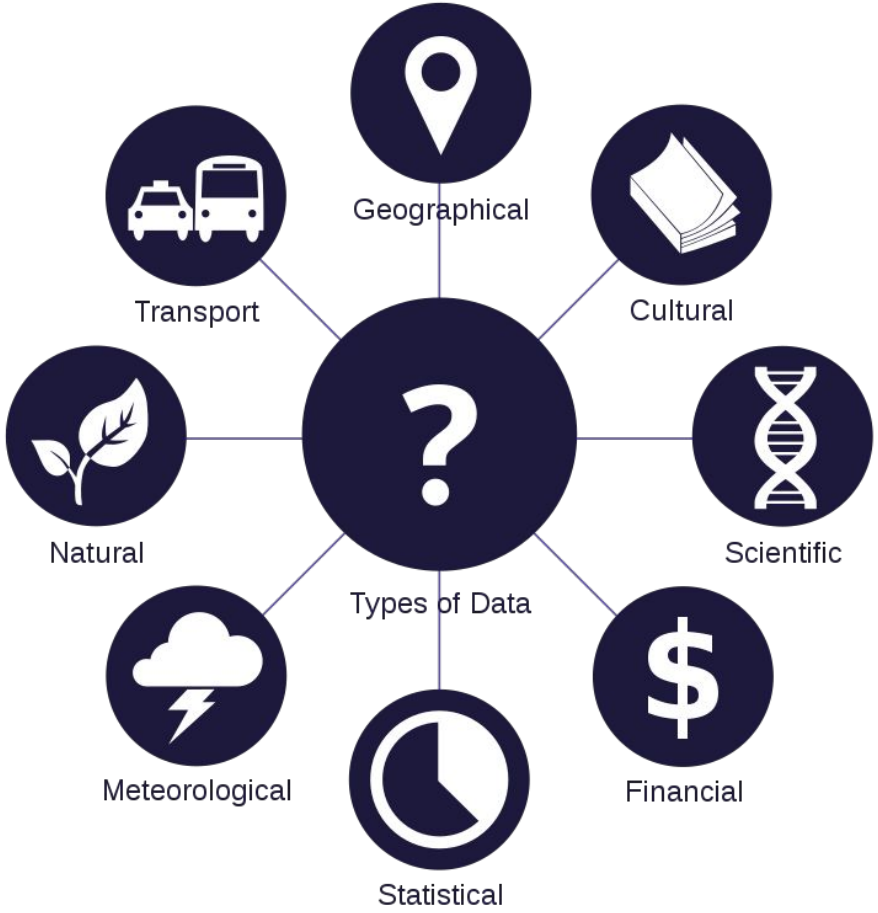
Machine Learning Models are only as Good as the Data They're Trained on

Deepchecks Community Blog | April 19, 2022

“A model is only as good as its data”

True or not,
data still play a fundamental role in ML and AI

Data



What types of data do you work with?

https://bit.ly/data_survey_01

What are Data?

Definition? [Cambridge Advanced Learner's Dictionary & Thesaurus]

- uncountable noun used with singular or plural verb
- "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer"

What are Data?

Definition? [Cambridge Advanced Learner's Dictionary & Thesaurus]

- uncountable noun used with singular or plural verb
- "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer"

What are Data?

Definition? [Cambridge Advanced Learner's Dictionary & Thesaurus]

- uncountable noun used with singular or plural verb
- "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer"

Data ≠ information

What are Data?

Definition? [Cambridge Advanced Learner's Dictionary & Thesaurus]

- uncountable noun used with singular or plural verb
- "information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer"

Data ≠ information

Data is potential information

What are Research Data?

Definition [DFG Guidelines (archived)]

Research data is an essential foundation for scientific work. The diversity of this data reflects the wide range of different scientific disciplines, research interests and research methods. Research data might include measurement data, laboratory values, audiovisual information, texts, survey data, objects from collections, or samples that were created, developed or evaluated during scientific work. Methodical forms of testing such as questionnaires, software and simulations may also produce important results for scientific research and should therefore also be categorised as research data.

Data in AI and ML

Types

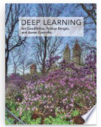
- Structured
  - numerical, categorical, time-series in spreadsheets or databases
  - metadata
- Unstructured
  - text, images, audio, video (can be structured, e.g. text with XML/TEI)
- Trained models (ready to be used)

From all disciplines [ÖFOS 2012]

- Natural sciences
- Technical sciences
- Human medicine, health sciences
- Agricultural sc., veterinary medicine
- Social sciences
- Humanities

Data and Metadata

- Metadata = “data about data”
- “Metadata is a statement about a potentially informative object”
  (Jeffrey Pomerantz, *Metadata*. The MIT Press Essential Knowledge series, p. 26)
- Metadata helps us find, access, understand, and reuse data
- No clear-cut boundary between data and metadata



Buch

Deep learning

Goodfellow, Ian, 1987- [VerfasserIn] ; Bengio, Yoshua [VerfasserIn] ; Courville, Aaron [VerfasserIn]
Cambridge, Massachusetts : London, England : The MIT Press ; [2016]

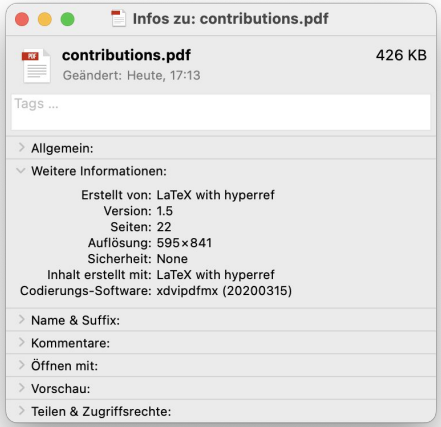
❝ *Deep Learning Ian Goodfellow Yoshua Bengio and Aaron Courville The MIT Press Cambridge Massachusetts London England Contents Website xiii Acknowledgments* ❞

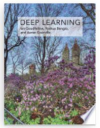
Verfügbar: FB Publizistik- und Kommunikationswissenschaft und Informatik / Signatur: Prof. Möller FG Visualization INF-7374 (mehrere Standorte) ›

Details

Titel	Deep learning
Verantwortliche	Ian Goodfellow, Yoshua Bengio and Aaron Courville
Ort/Verlag	Cambridge, Massachusetts : London, England : The MIT Press
Erscheinungsjahr	[2016]
Umfang/Format	xxii, 775 Seiten, Illustrationen, Diagramme
Standardnummern	ISBN: 9780262035613
Beschreibung	Literaturverzeichnis: Seite [711]-766
Verknüpfte Titel	Adaptive computation and machine learning
Sprache	Englisch
Person/Institution	Goodfellow, Ian, 1987- [VerfasserIn] › Bengio, Yoshua [VerfasserIn] › Courville, Aaron [VerfasserIn] ›
Basisklassifikation	54.72 - Künstliche Intelligenz ›
Regensburger Verbundklassifkation	ST 302 › ST 301 › ST 300 ›
Schlagwörter	Maschinelles Lernen › Künstliche Intelligenz › Deep learning ›
Quelle	UB Wien
Permalink	https://ubdata.univie.ac.at/AC13299661

https://ubdata.univie.ac.at/AC13299661





Buch

Deep learning

Goodfellow, Ian, 1987- [VerfasserIn] ; Bengio, Yoshua [VerfasserIn] ; Courville, Aaron [VerfasserIn]

Cambridge, Massachusetts : London, England : The MIT Press ; [2016]

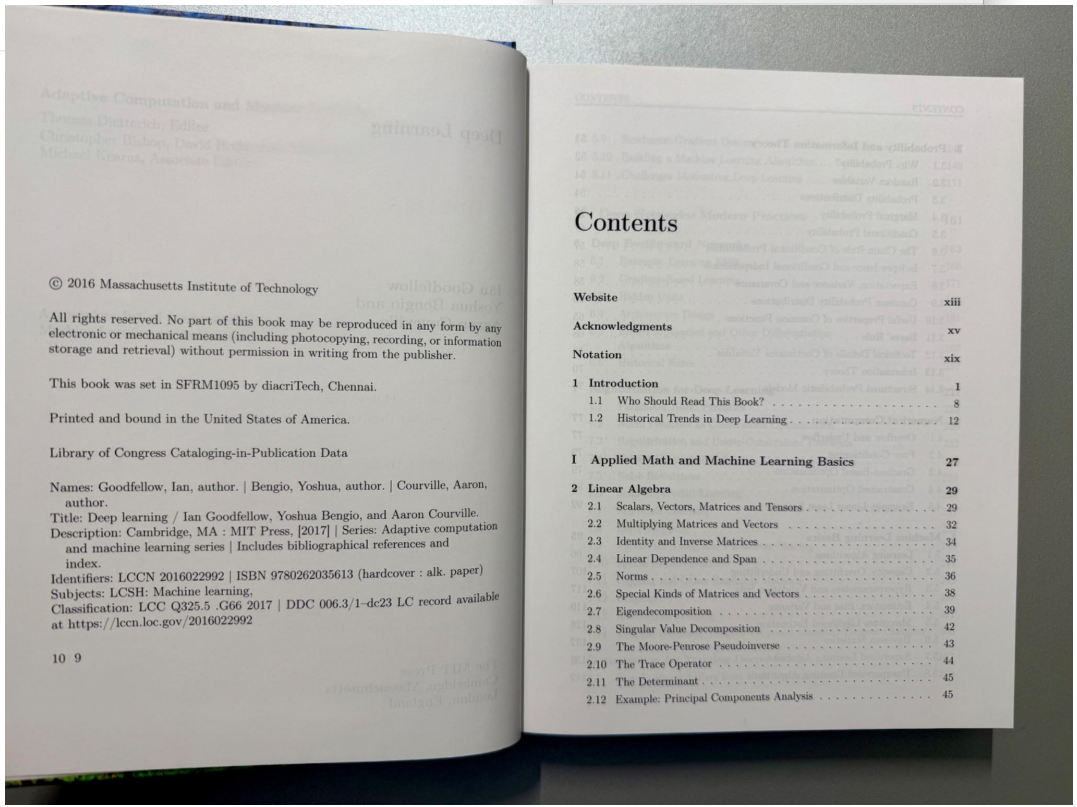
❝ *Deep Learning Ian Goodfellow Yoshua Bengio and Aaron Courville The MIT Press Cambridge Massachusetts London England Contents Website xiii Acknowledgments* ❞

Verfügbar: FB Publizistik- und Kommunikationswissenschaft und Informatik / Signatur: Prof. Möller FG Visualization INF-7374 (mehrere Standorte) ›

Details

Titel	Deep learning
Verantwortliche	Ian Goodfellow, Yoshua Bengio and Aaron Courville
Ort/Verlag	Cambridge, Massachusetts : London, England : The MIT Press
Erscheinungsjahr	[2016]
Umfang/Format	xxii, 775 Seiten, Illustrationen, Diagramme
Standardnummern	ISBN: 9780262035613
Beschreibung	Literaturverzeichnis: Seite [711]-766
Verknüpfte Titel	Adaptive computation and machine learning
Sprache	Englisch
Person/Institution	Goodfellow, Ian, 1987- [VerfasserIn] › Bengio, Yoshua [VerfasserIn] › Courville, Aaron [VerfasserIn] ›
Basisklassifikation	54.72 - Künstliche Intelligenz ›
Regensburger Verbundklassifkation	ST 302 › ST 301 › ST 300 ›
Schlagwörter	Maschinelles Lernen › Künstliche Intelligenz › Deep learning ›
Quelle	UB Wien
Permalink	https://ubdata.univie.ac.at/AC13299661

https://ubdata.univie.ac.at/AC13299661



This book was set in SFRM1095 by diacriTech, Chennai.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Goodfellow, Ian, author. | Bengio, Yoshua, author. | Courville, Aaron, author.
Title: Deep learning / Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Description: Cambridge, MA : MIT Press, [2017] | Series: Adaptive computation and machine learning series | Includes bibliographical references and index.
Identifiers: LCCN 2016022992 | ISBN 9780262035613 (hardcover : alk. paper)
Subjects: LCSH: Machine learning,
Classification: LCC Q325.5 .G66 2017 | DDC 006.3/1–dc23 LC record available at https://lccn.loc.gov/2016022992

10 9

Contents

Website xiii

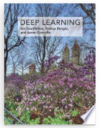
Acknowledgments xv

Notation xix

1 Introduction 1
1.1 Who Should Read This Book? 8
1.2 Historical Trends in Deep Learning 12

I Applied Math and Machine Learning Basics 27

2 Linear Algebra 29
2.1 Scalars, Vectors, Matrices and Tensors 29
2.2 Multiplying Matrices and Vectors 32
2.3 Identity and Inverse Matrices 34
2.4 Linear Dependence and Span 35
2.5 Norms 36
2.6 Special Kinds of Matrices and Vectors 38
2.7 Eigendecomposition 39
2.8 Singular Value Decomposition 42
2.9 The Moore-Penrose Pseudoinverse 43
2.10 The Trace Operator 44
2.11 The Determinant 45
2.12 Example: Principal Components Analysis 45



Buch

Deep learning

Goodfellow, Ian, 1987- [VerfasserIn] ; Bengio, Yoshua [VerfasserIn] ; Courville, Aaron [VerfasserIn]
Cambridge, Massachusetts : London, England : The MIT Press ; [2016]

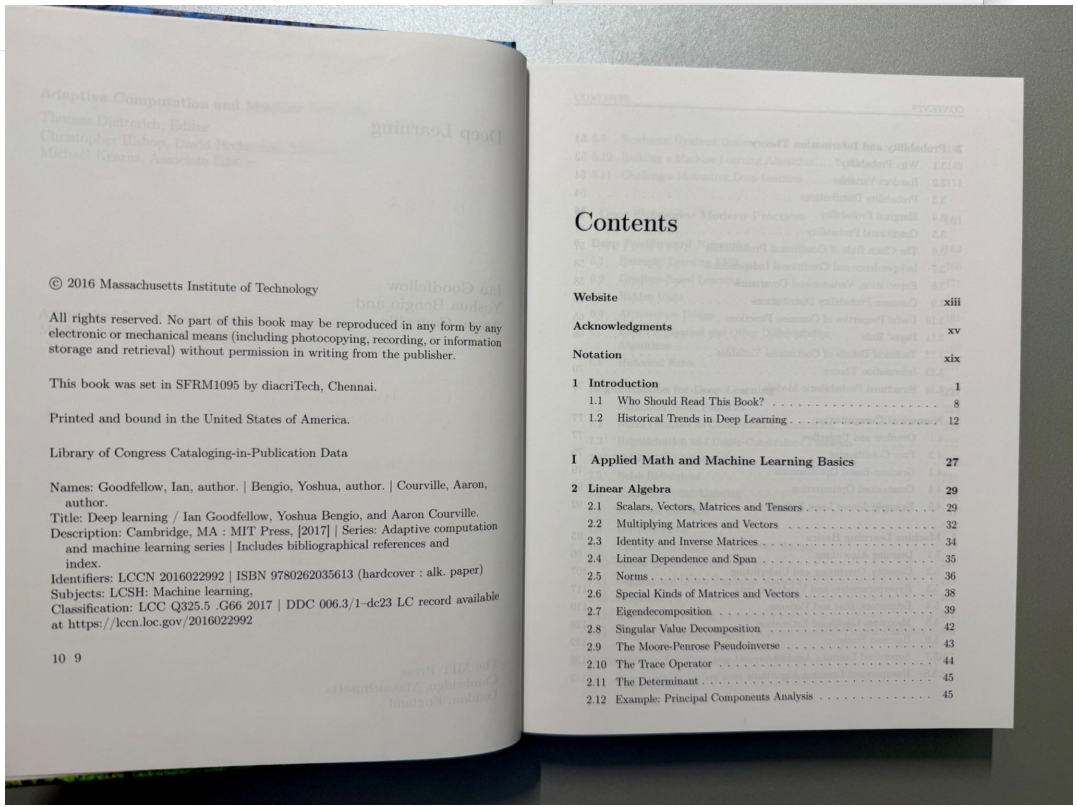
❝ *Deep Learning Ian Goodfellow Yoshua Bengio and Aaron Courville The MIT Press Cambridge Massachusetts London England Contents Website xiii Acknowledgments* ❞

Verfügbar: FB Publizistik- und Kommunikationswissenschaft und Informatik / Signatur: Prof. Möller FG Visualization INF-7374 (mehrere Standorte) ›

Details

Titel	Deep learning
Verantwortliche	Ian Goodfellow, Yoshua Bengio and Aaron Courville
Ort/Verlag	Cambridge, Massachusetts : London, England : The MIT Press
Erscheinungsjahr	[2016]
Umfang/Format	xxii, 775 Seiten, Illustrationen, Diagramme
Standardnummern	ISBN: 9780262035613
Beschreibung	Literaturverzeichnis: Seite [711]-766
Verknüpfte Titel	Adaptive computation and machine learning
Sprache	Englisch
Person/Institution	Goodfellow, Ian, 1987- [VerfasserIn] › Bengio, Yoshua [VerfasserIn] › Courville, Aaron [VerfasserIn] ›
Basisklassifikation	54.72 - Künstliche Intelligenz ›
Regensburger Verbundklassifkation	ST 302 › ST 301 › ST 300 ›
Schlagwörter	Maschinelles Lernen › Künstliche Intelligenz › Deep learning ›
Quelle	UB Wien
Permalink	https://ubdata.univie.ac.at/AC13299661

https://ubdata.univie.ac.at/AC13299661



This book was set in SFRM1095 by diacriTech, Chennai.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Goodfellow, Ian, author. | Bengio, Yoshua, author. | Courville, Aaron, author.
Title: Deep learning / Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Description: Cambridge, MA : MIT Press, [2017] | Series: Adaptive computation and machine learning series | Includes bibliographical references and index.
Identifiers: LCCN 2016022992 | ISBN 9780262035613 (hardcover : alk. paper)
Subjects: LCSH: Machine learning,
Classification: LCC Q325.5 .G66 2017 | DDC 006.3/1–dc23 LC record available at https://lccn.loc.gov/2016022992

10 9

Contents

Website xiii

Acknowledgments xv

Notation xix

1 Introduction 1
1.1 Who Should Read This Book? 8
1.2 Historical Trends in Deep Learning 12

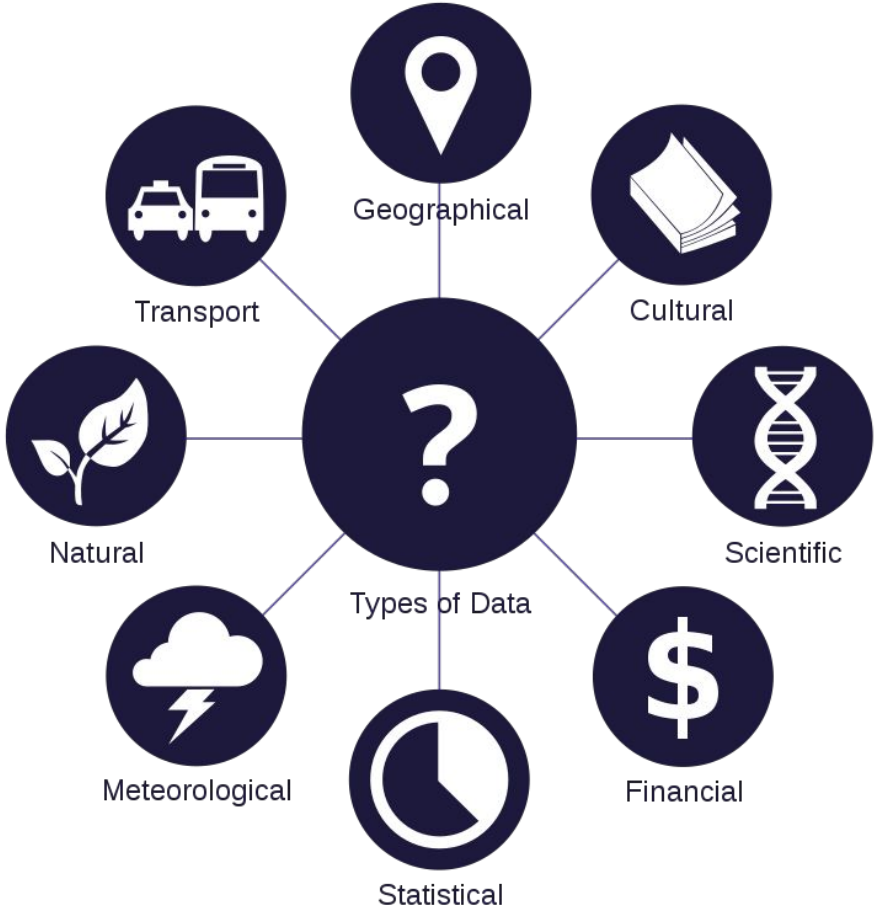
I Applied Math and Machine Learning Basics 27

2 Linear Algebra 29
2.1 Scalars, Vectors, Matrices and Tensors 29
2.2 Multiplying Matrices and Vectors 32
2.3 Identity and Inverse Matrices 34
2.4 Linear Dependence and Span 35
2.5 Norms 36
2.6 Special Kinds of Matrices and Vectors 38
2.7 Eigendecomposition 39
2.8 Singular Value Decomposition 42
2.9 The Moore-Penrose Pseudoinverse 43
2.10 The Trace Operator 44
2.11 The Determinant 45
2.12 Example: Principal Components Analysis 45

More potential data sources

- Metadata are also data
- Scientific or grey literature in unstructured PDFs
- Badly (or not yet) OCRed printed sources
- Printed – and not yet digitized – documents
  - Often containing structured data, but in printed form (e.g. tables)
- Other materials in libraries, museums, archives, and other Cultural Heritage institutions
- New data can still be created (in several ways) – and is being created
- The question of "peak data"
  - "We've achieved peak data and there'll be no more" (I. Sutskever, co-founder of OpenAI)
  - One interesting post by S. Majstorovic (from which some of the above examples come)

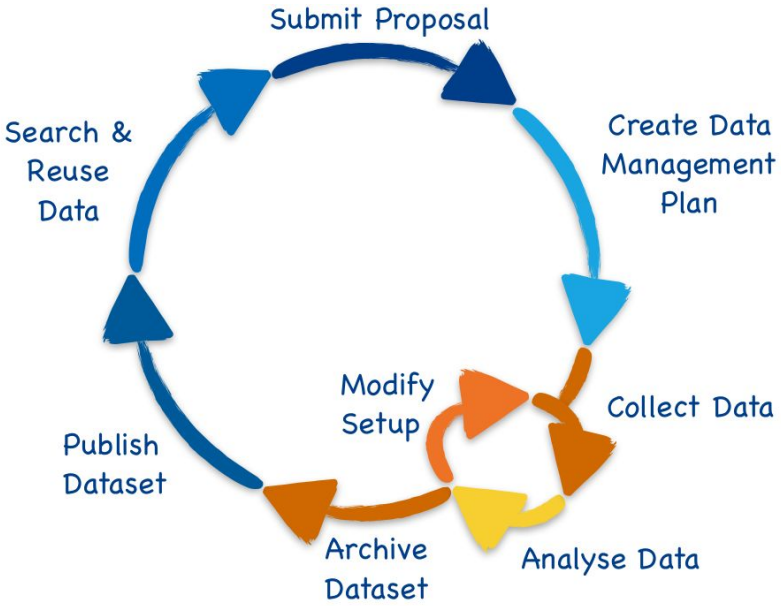
Research Data



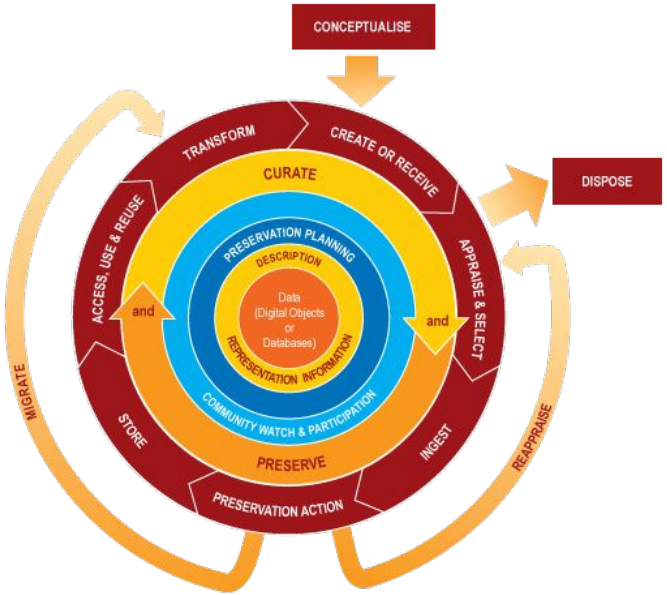
Data Lifecycle

[IANUS Lebenszyklus] [UK Data Service]

Arbitrarily Complex Data Lifecycles

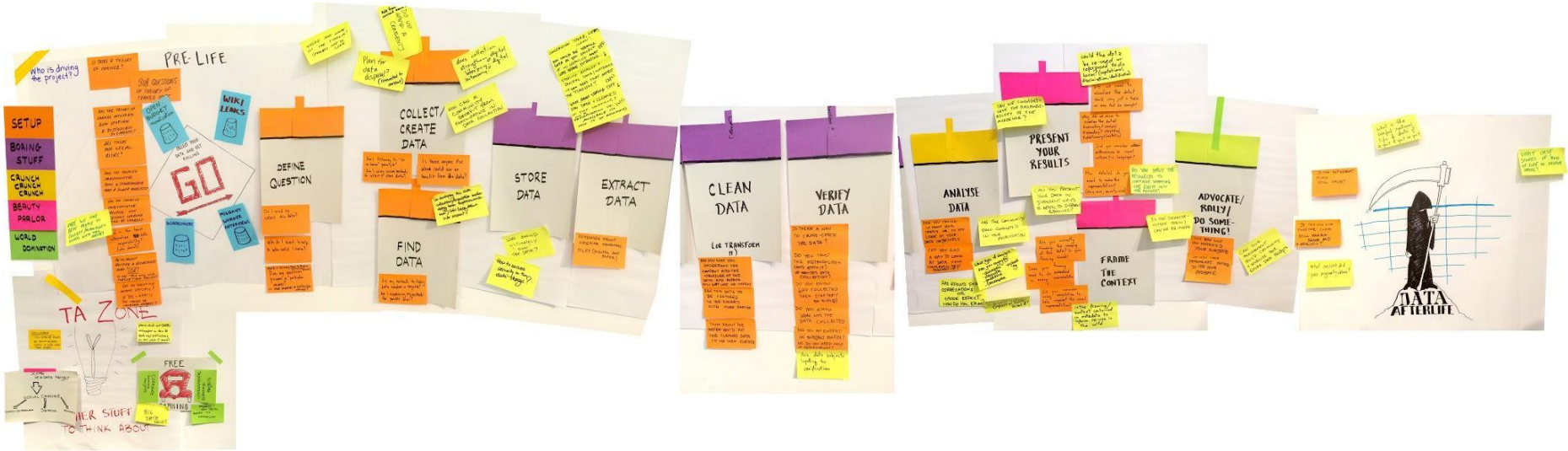


HZDR, Oliver Knodel, Datenlebenszyklus



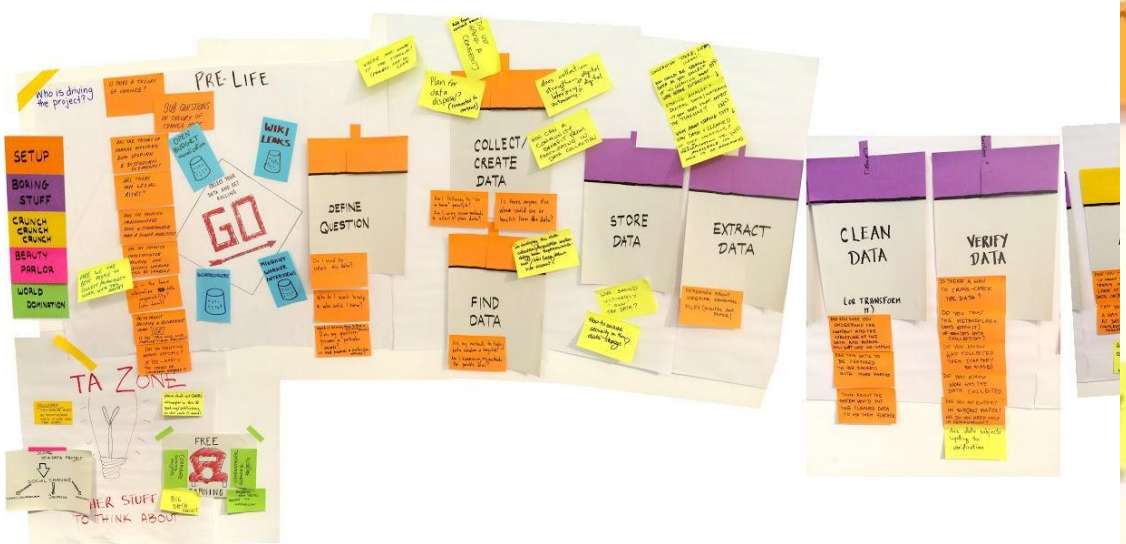
DCC, Curation Lifecycle Model

Arbitrarily Complex Data Lifecycles



Mushonz, CC BY-SA 4.0, via Wikimedia Commons

Arbitrarily Complex Data Lifecycles



Data Management

Why Data Management?

BASIC DATA

Knowledge at the starting point of a research project is based on data.

Why Data Management?

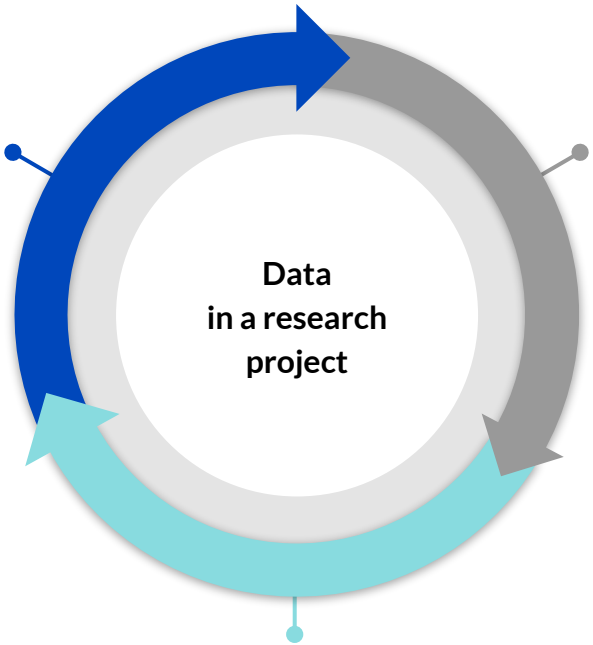
BASIC DATA

Knowledge at the starting point of a research project is based on data.

NEW DATA

Processing and analysing data creates new data.

Why Data Management?



RESEARCH RESULT

New knowledge based on new data.

BASIC DATA

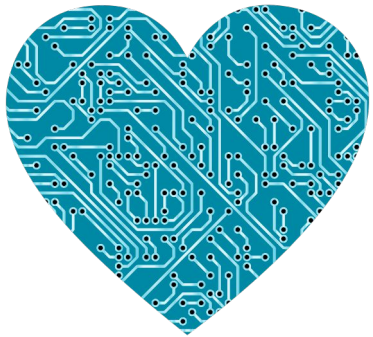
Knowledge at the starting point of a research project is based on data.

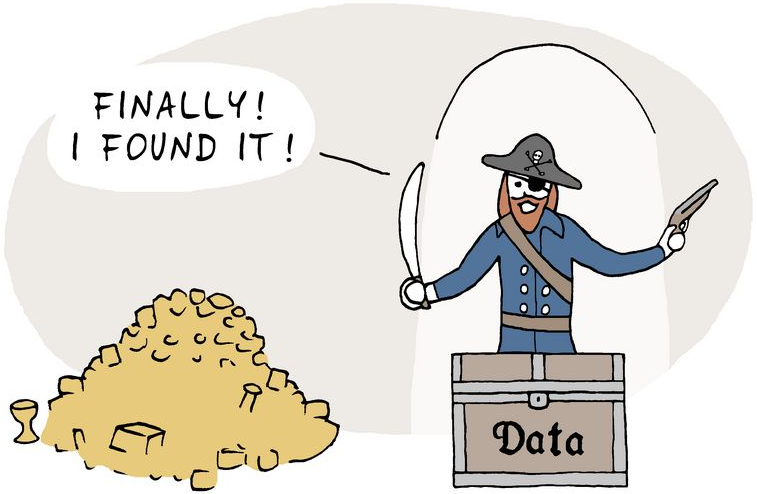
NEW DATA

Processing and analysing data creates new data.

Why Data Management?

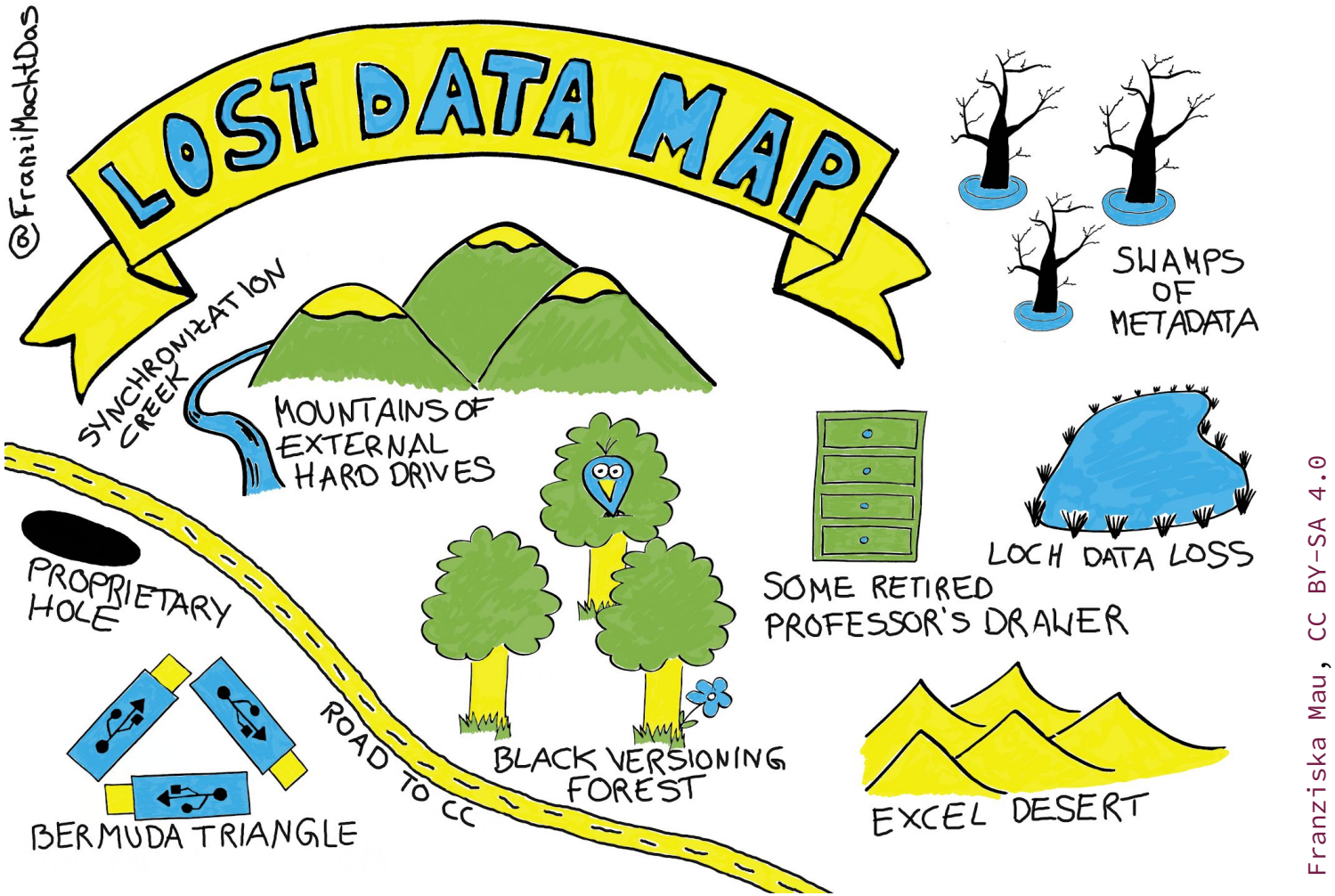
Data are the heart of a project.
Data should be cared for.
Keep the data lifecycle going.





Dataedo /cartoon

Piotr@Dataedo



Why Data Management?

- Data form the starting point of a research project
- Knowledge about a topic is based on data
- New data emerge in the course of a project
- Processing and analysing data leads to new knowledge
- Research results are based on data

Plan

Data Management Plan (DMP)

«A data management plan or DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed.» [Wikipedia]

[DCC, Data Management Plans]
[OpenAIRE & EUDAT, ARGOS]
[DMPonline]
[Science Europe]
[Forschungsdaten.info, Der Datenmanagementplan]
[FWF, Forschungsdatenmanagement]
[IANUS, Datenmanagement]

Data Management Plan (DMP)

Project specifics

What and how?
Data types & volume
Costs

Data Management Plan (DMP)

Data Management Plan (DMP)

Data Management Plan (DMP)

[EU AI standards]
[NIST standards]
[Austrian Standards on AI]

[FAIR Data Principles]
[GO FAIR]
[Jones & Grootveld 2017]
[FOSTER Open Science]

Data Management Plan in practice

Data Management Plan Example: FWF DMP

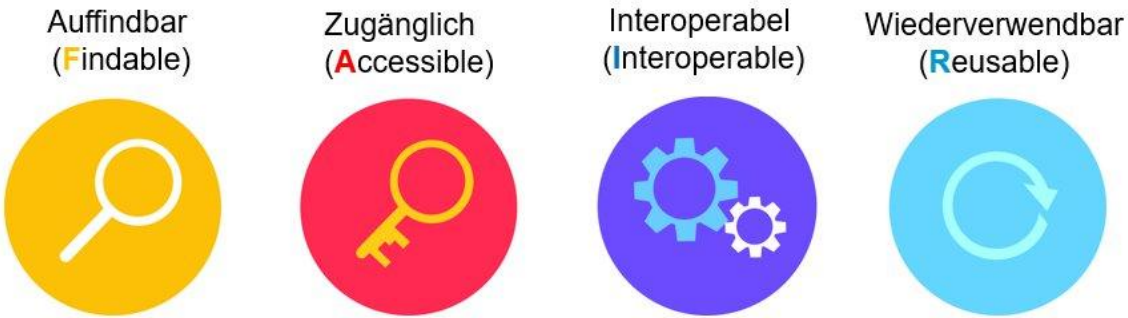
Data Management Tools:

- TU Wien: DAMAP
- DMP tools list, HOW-TO link

DMP should be revised and updated during the project!

FAIR data

FAIR data principles



[**FAIR Data Principles**], [**Wilkinson et al. 2016**] [**GO FAIR**] [Forschungsdaten.info, FAIRE Daten] [Top 10 FAIR Data & Software Things] [A FAIRy tale] [FAIR Data Austria] [Jones & Grootveld 2017] [FAIR for AI]

FAIR data principles

Findable

- metadata & persistent identifier

Accessible

- standardised protocol (machine-readable)

Interoperable

- standards

Reusable

- documentation, standards, license

Digital archives

take care of most of it.

Core Task Areas of Data Management

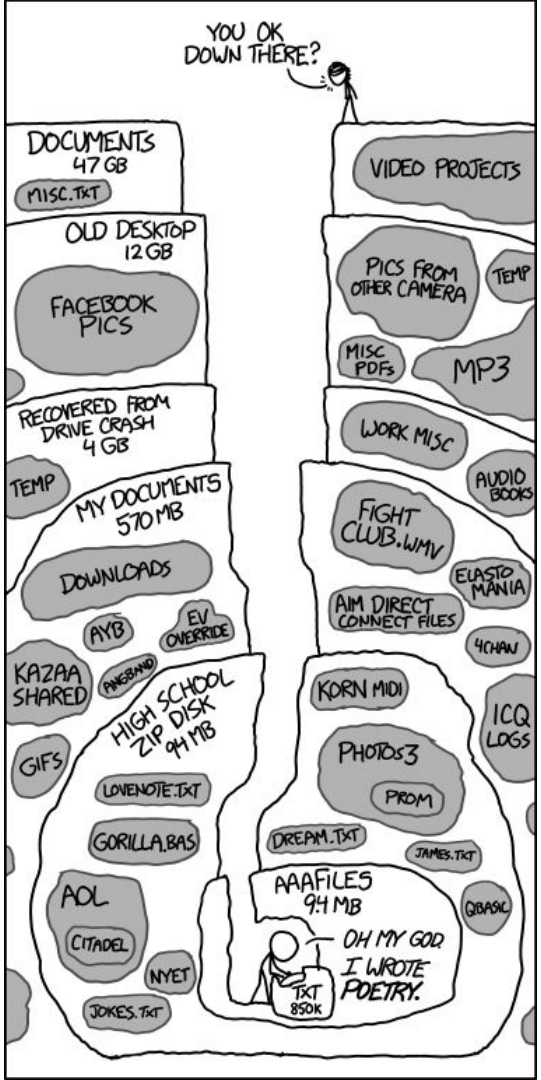
Core Task Areas

- Organise & structure
  - directory structure
  - file naming
- Versioning
- File formats
- Document
  - unstructured but extensive
  - structured documentation with metadata
  - metadata standards and controlled vocabularies
- Rights & Licensing
- Disseminate & preserve

Organise & Structure

Why Organise Files and Folders?

- Increase efficiency
- Help others in finding & understanding (including especially future you)



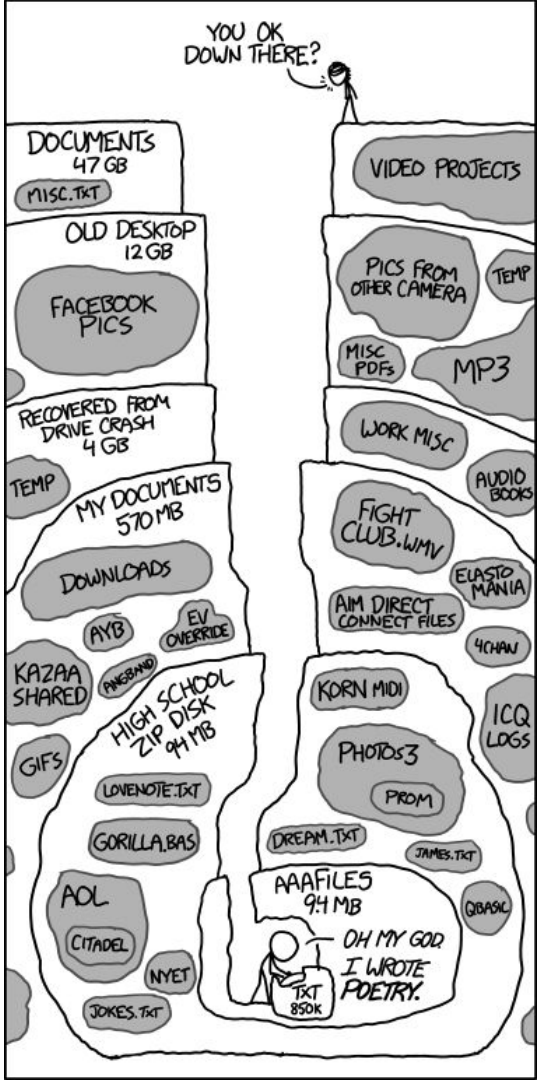
Old Files, XKCD 1360, CC BY-NC 2.5

Why Organise Files and Folders?

- Increase efficiency
- Help others in finding & understanding (including especially future you)

➤ **Two parts:**
File and folder names
Folder structure

[CESSDA Training]
[Stanford Libraries]
[IANUS Dateiverwaltung]



Best Practice: File and Folder Names

Good

- Use unambiguous and unique names
- Use descriptive names
- Be consistent
- As long as necessary,
  as short as possible

Best Practice: File and Folder Names

Better

- Use date in ISO format YYYY-MM-DD

```
2021-11-01_presentation.pdf
```

- Indicate versioning

```
DMP_v1-0.docx
DMP_v2-0.docx
```

- Use leading zeros

```
DMP_v2-1.docx
```

- Do not use dots in file name;
  this is reserved for file extension

```
001_image.png
002_image.png
```

Best Practice: File and Folder Names

Best

- Do not use spaces
- Stay alphanumeric (letters from Latin alphabet and numbers)
- Do not use any special characters, diacritics, accents etc. (e.g. `? / ö ẑ ¢`)
- Use hyphen (-) and underscore (_) for element separation
- Write down naming conventions
  → more in the section “Document”

```
data management ->
hyphen                  data-management
underscore              data_management
join                    datamanagement
camelCase               dataManagement

Übergröße ->    Uebergroesse
                Uebergroesze

2021-11-01_presentation.pdf
```

Popular Naming Conventions

- PascalCase
  - `DateOfBirth`
- camelCase
  - `dateOfBirth`
- snake_case
  - `date_of_birth`
- kebab-case
  - `date-of-birth`

For more info, Camel case (Wikipedia)

Emoji One, CC BY-SA 4.0, via Wikimedia Commons

Naming recommendations

Valid not only for

- **File and folder names**

But also…

- **Variables**
- **Functions**
- **Column headers**
- etc.

Source: Name It Like You Mean It: Variable Names For Well-Designed Survey Research

Name	Label
e3	Have you bought a...
m_e3_0	No vehicles
m_e3_1	Bicycle
m_e3_2	Motorbike
m_e3_3	Tuk-tuk?
m_e3_4	Other vehicle?

… and remember to document abbreviations / conventions!

File and Folder Structure

5S Method [Fuchs 2019]

- Sort
- Set in order
- Shine
- Standardise
- Sustain

File and Folder Structure

5S Method [Fuchs 2019]

- Sort
- Set in order
- Shine
- Standardise
- Sustain

- Check files & folders; remove unnecessary; attic area; productive

File and Folder Structure

5S Method [Fuchs 2019]

- Sort
- Set in order
- Shine
- Standardise
- Sustain

- Check files & folders; remove unnecessary; attic area; productive
- Arrange items in useful way; apply naming conventions

File and Folder Structure

5S Method [Fuchs 2019]

- Sort
- Set in order
- Shine
- Standardise
- Sustain

- Check files & folders; remove unnecessary; attic area; productive
- Arrange items in useful way; apply naming conventions
- Regularly inspect and clean

File and Folder Structure

5S Method [Fuchs 2019]

- Sort
- Set in order
- Shine
- Standardise
- Sustain

- Check files & folders; remove unnecessary; attic area; productive
- Arrange items in useful way; apply naming conventions
- Regularly inspect and clean

- Write down rules & conventions; group responsibility

File and Folder Structure

5S Method [Fuchs 2019]

- Sort
- Set in order
- Shine
- Standardise
- Sustain

- Check files & folders; remove unnecessary; attic area; productive
- Arrange items in useful way; apply naming conventions
- Regularly inspect and clean

- Write down rules & conventions; group responsibility
- Make it a habit; train group; protocol changes when needed

Folder Structuring Pointers [IANUS Dateiverwaltung]

Depends on

- Project (type, size, work packages)
- Data (collections, types, processing state)
- Activities (work packages, methods)
- Documentation & administration files
- External guidelines

Folder Structuring Pointers [IANUS Dateiverwaltung]

Depends on

- Project (type, size, work packages)
- Data (collections, types, processing state)
- Activities (work packages, methods)
- Documentation & administration files
- External guidelines

Consider

- Hierarchy (flat vs. deep)
- Understandability
- Naming best practice

Avoid

- File duplication
- File shortcuts

File checker

https://github.com/acdh-oeaw/repo-file-checker

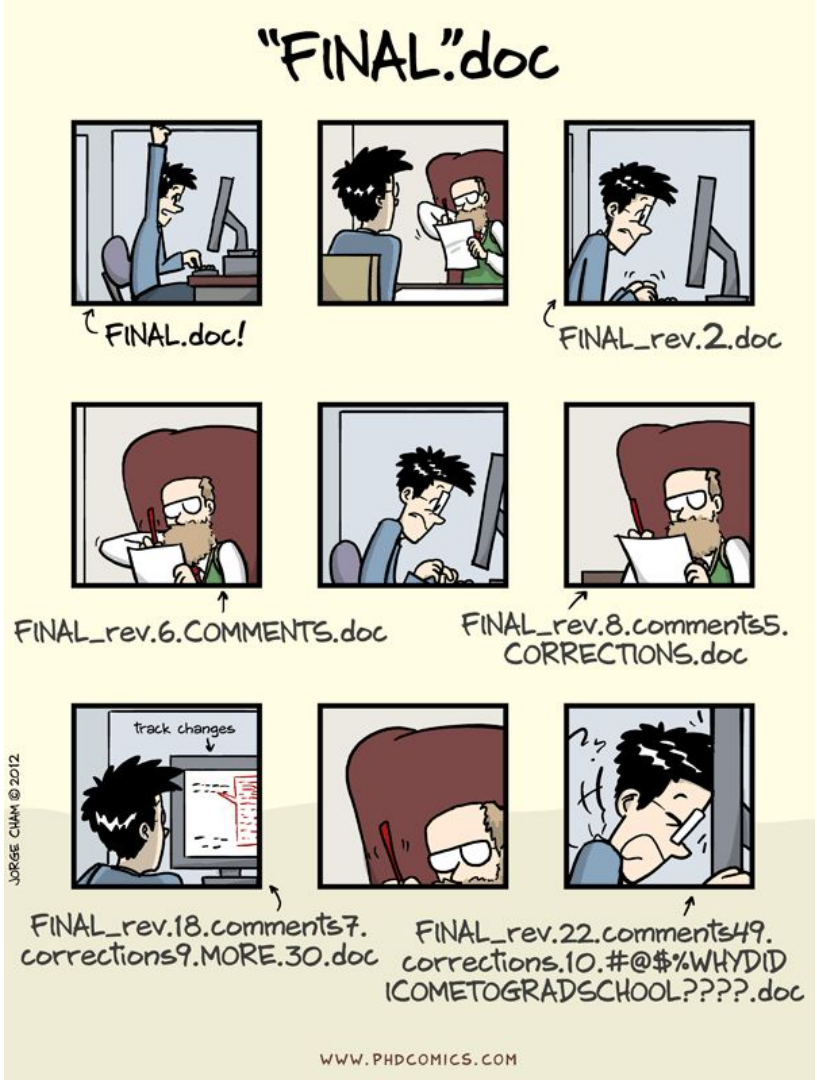
- Tool developed in the context of ARCHE
- Outputs to JSON but can also create reports in HTML and CSV
- Checks duplicate files based on hash
- Invalid filenames
- Corrupt image files
- MIME/extension mismatch
- Accepted formats (ARCHE-specific)
- Runs a series of other checks
  - PDF/A
  - Empty directories
  - Password protected files
  - Schema validation for XML files
  - …

severity	error	count
ERROR	Duplicated file	146472
ERROR	Not a PDF/A (in practice less because it's few errors per each file)	125404
ERROR	Format not accepted	54963
ERROR	Invalid filename	56465
ERROR	Unknown MIME	12878
ERROR	MIME/extension mismatch	12855
ERROR	Empty directory	226
ERROR	Corrupted image	17
ERROR	RuntimeException (local charset used in filename inside a ZIP archive)	9
ERROR	Password protected file	8
ERROR	File contains Byte Order Mark	1
ERROR	XML (missing doctype declaration)	1
WARNING	Rotated image	9507
WARNING	Format not preferred	5125
WARNING	XML (missing schema and/or encoding declaration)	1163

Versioning

Versioning: Why?

- Collaborative work
- Keep track of changes
- Support experimentation and reproducibility of research results
- Allow going back to previous states
- Adapt to different scenarios
- Simplify reuse by the community



Versioning: How [Stanford Libraries; IANUS Versionskontrolle]

Manually

- Indicate version in file name
  - date
  - version number
  - labels like ‘draft’ or ‘final’
- Include information in file
  - basic information (when, who)
  - changelog
- Keep changelog in separate file
- [semver], [semverdoc - MMVC], [semver for AI]

Automated through software

(especially for groups)

- Automatically tracks changes and stores versions
- basic: cloud service, e.g. ownCloud or Google Drive
- advanced: version-control software, e.g. Git or Subversion
  - (Git or Git-like version control also in repositories for ML models, like HuggingFace)

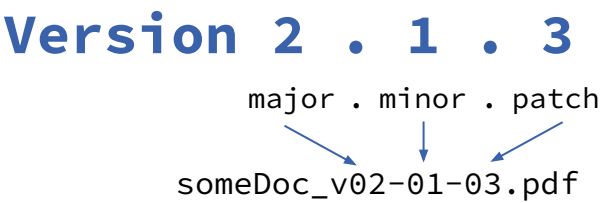
Semantic Versioning

Software [semver]

- Major: incompatible API changes
- Minor: new functionality in a backward compatible manner
- Patch: backward compatible bug fixes

Documents [semverdoc]

- Major: significant changes
- Minor: information added or removed
- Patch: minor changes, e.g. fixing typos



Semantic Versioning for AI Models

Many components at play when training an AI model:

- Data (and its metadata)
- Settings/configuration (including hyperparameters)
- Pre-trained artifacts (word embeddings, foundational models (e.g. BERT), etc.)
- Model
- Evaluation metrics
- Code (with dependencies)

Single components ≠ their combination

Semantic Versioning for AI Models

A possible solution, based on [semver for AI 1.0.0]:

- Version the single components/artifacts (or get version info from source)
- Version the combination of all these components

Examples:

Supervised model

We add more examples of a specific category

`dataset 1.0.0 → 1.1.0`

We improve eval metric by 2%

`model 1.0.0 → 1.1.0`

`eval 1.0.0 → 1.1.0`

`1.0.0 → 1.1.0`

Unsupervised model

We change the number of clusters

`config 1.0.0 → 1.1.0`

We improve eval metric by 5%

`model 1.0.0 → 2.0.0`

`eval 1.0.0 → 2.0.0`

`1.0.0 → 2.0.0`

Criteria for Semantic Versioning in AI

Datasets [semver for AI]

- Major: changes in the schema or semantics, instances removed/changed
- Minor: new instances that alter distribution of dataset
- Patch: new instances that do not alter distribution

Models [semver for AI]

- Major: add/remove hyperparameters, eval metrics change +/-3%, …
- Minor: change hyperparameter, eval metrics change between +/-1% and 3%, …
- Patch: change random seed, eval metrics change less than +/-1%

Criteria for Semantic Versioning in AI

- Different criteria are possible
  - Difficult to establish when an AI model *changes* significantly
  - Consider the broader scenario in which this AI model is used
  - Change in type of architecture → New version or completely new model?
- Version number + additional labels (before/after)
  - `cat-detector_cnn_v3-0-0`
  - `cat-detector_transf_v1-0-0`
  - `cat-detector_v3-1-2-alpha`
- Calendar versioning for specific artifacts
  - `cat-dataset_2024-12-31`
  - `cat-dataset_2025-01-19`

OpenAI Versioning Conventions

MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
gpt-4o ↳ gpt-4o-2024-08-06	128,000 tokens	16,384 tokens
gpt-4o-2024-11-20	128,000 tokens	16,384 tokens
gpt-4o-2024-08-06	128,000 tokens	16,384 tokens
gpt-4o-2024-05-13	128,000 tokens	4,096 tokens
chatgpt-4o-latest ↳ GPT-4o used in ChatGPT	128,000 tokens	16,384 tokens

[OpenAI Docs - Models]

OpenAI Versioning Conventions

MODEL
gpt-4o ↳ gpt-4o-2024-08-06
gpt-4o-2024-11-20
gpt-4o-2024-08-06
gpt-4o-2024-05-13
chatgpt-4o-latest ↳ GPT-4o used in ChatGPT

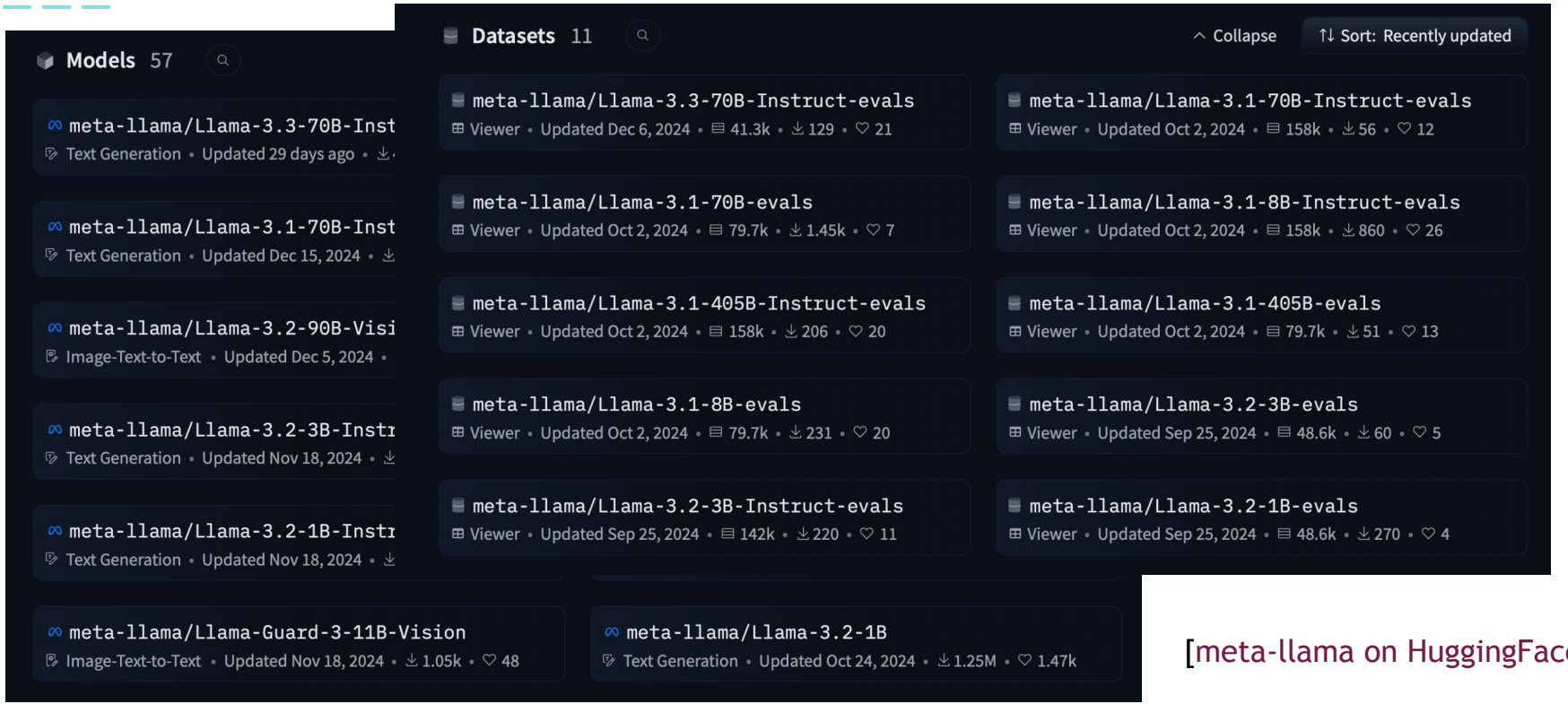
MODEL	CONTEXT WINDOW	MAX OUTPUT TOKENS
gpt-4o-realtime-preview ↳ gpt-4o-realtime-preview-2024-12-17	128,000 tokens	4,096 tokens
gpt-4o-realtime-preview-2024-12-17	128,000 tokens	4,096 tokens
gpt-4o-realtime-preview-2024-10-01	128,000 tokens	4,096 tokens
gpt-4o-mini-realtime-preview ↳ gpt-4o-mini-realtime-preview-2024-12-17	128,000 tokens	4,096 tokens
gpt-4o-mini-realtime-preview-2024-12-17	128,000 tokens	4,096 tokens

[OpenAI Docs - Models]

Llama Versioning Conventions

[meta-llama on HuggingFace]

Llama Versioning Conventions



[meta-llama on HuggingFace]

Using Git(Hub) for Versioning AI Models

- One repo for **each model**
- Organize your **repo structure**
- Keep **data** separate from code
- **Branches** for fixes/experiments
- Use **commit hashes** and/or **version numbers**
  - e.g. version numbers only for tags/releases
- Use Git LFS for trained models
- Automate with GitHub Actions

```
├── data/
│   ├── raw/ # Raw data (not tracked by Git)
│   ├── processed/ # Processed data (optional tracking)
├── models/ # Trained models (use Git LFS or DVC)
├── notebooks/ # Jupyter notebooks for experimentation
├── src/ # Source code for training and evaluation
├── requirements.txt # Package dependencies
└── config.yaml # Configuration file for experiments
```

Example of a Git repo structure from R. Rathore, Git and AI: Using Git for Machine Learning Models Versioning

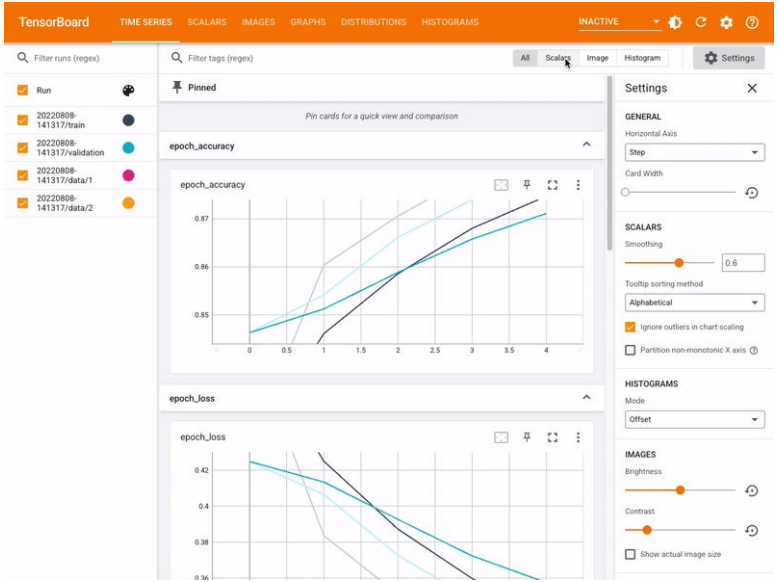
Tools for Tracking the ML Lifecycle

- MLflow [Getting started]
- Weights & Biases [Quickstart]
- Neptune [Quickstart, Tutorial]
- TensorBoard [Get started]

Especially for data:

- DVC [Get started]
- LakeFS [Quickstart]
- Pachyderm [Get started]

Other tools in Document section



From TensorBoard website

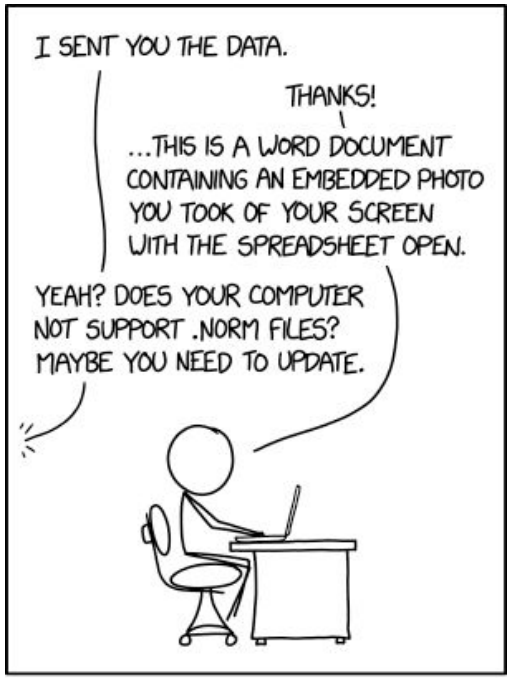
File Formats




File Formats for Sharing and Preservation

Principles [IANUS Dateiformate]

- Widely in use
- Non-proprietary
- Open standard
- Uncompressed or lossless compression
- Unencrypted
- Consider requirements and standards



.NORM Normal File Format, XKCD 2116, CC BY-NC 2.5

File Format Choices [ARCHE Formats]

Formatted Text

- PDF/A-1, PDF/A-2
- ~~PDF, doc~~

Structured Text

- xml, html, txt, md, tex (UTF-8, no BOM)
- ~~PDF, docx, doc, odt, indd~~

Plain Text

- txt (UTF-8, no BOM)
- ~~PDF, docx, odt~~

Images

- tiff, dng
- ~~jpeg, psd, gif~~

Vector Images

- svg
- ~~ai, indd, ps, dwg, dxf~~

Spreadsheets, Tables

- csv (UTF-8, no BOM)
- ~~xls, ods~~

(Copy)Rights & Licensing

Legal framework and good practice

01 What data can I use?

- IPR (copyright), neighbouring/related rights
- other restrictions for publishing data (property rights)
- ethical consideration, GDPR
- good scientific practice (citation)

- Are you allowed to use existing data?
- How can you use the existing data? E.g. are there requirements and restrictions for publishing the data?
- Will others be allowed to use your data? How?

➢ **Observe applicable legislation, contracts and policies.**

Copyright (Urheberrecht)

- applies to the rights on **original creative work**, the “literary and artistic work”
- creator is copyright holder

Definition [Berne Convention]

The expression ‘literary and artistic works’ shall include every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression.

Austrian law: “Urheber” & “Urheberrechtsgesetz (UrhG)”
[Understanding Copyright and Related Rights (WIPO 2016)]
[OpenAIRE Toolkit for Researchers on Legal Issues]
[Europeana Copyright Community in Europeana Pro network]

Copyright (Urheberrecht)

- There is no universal Copyright Law
- Copyright Laws are territorial, but there are attempts to reach common grounds - i.e. Conventions (like **Berne Convention**)
- Austrian Copyright Law:
  - Moral rights
  - Economic rights
  - Exceptions are written in the Law (no Fair Use in Austria!)
  - Automatically from the creation of the work until 70 years after death (the rights transfer automatically to the descendants)
  - Related rights from creation/publication of the work for 50 or 70 years

Public Domain

- Public domain: no more copyright (expired or waived)
- Digitised public domain works should remain public, however, other restrictions can apply!

➢ **Public domain works can be used, copied, remixed, etc. without any restrictions**

➢ **Even if a work is public domain, the good scientific practice requires citation**

- [Public domain calculators by IP Squared 2011]
- [Europeana Public Domain Charter 2010]

Permissible / open licenses

- License is a legally binding statement expressing conditions of how a work can be (re)-used.
- Check what a licenses allows for (publishing, remixing, derivatives etc.)

➢ It is a good idea to keep records of access rights, licenses, and acknowledgments to demonstrate compliance.

No license / statement, what now?

- Exceptions in Austrian Copyright Law **(§ 42 UrhG)**:
  - **Private & Research Use**
  - **Quotations**
  - **Educational Use**
  - **Text and Data Mining**
- **Non-commercial, attribution, no derivatives.**
- For anything else, you have to seek permission from the rights holder.

➤ **It is a good idea to keep records of access rights, licenses, and acknowledgments to demonstrate compliance.**

Related rights & databases

- Facts (measurements, counts…) and ideas are not copyrightable.
- Databases can be protected by copyright/related rights:
  - **Database Copyright (§ 40f UrhG)**: Protects databases as intellectual creations if they are the author's own intellectual work due to the selection or arrangement of their content.
  - **Database Sui Generis Right (§ 76c UrhG)**: Protects databases if substantial investments (financial, labor, or other resources) have been made to obtain, verify, or present their contents, even if the database lacks originality.

Contractual restrictions & property rights

- Contractual restrictions only apply to the parties signing the contract.
- Sometimes owners of works enforce restrictions on works that are public domain.

Ethical considerations & GDPR

- Using personal or sensitive data for machine learning is tightly regulated and requires adherence to various legal and ethical principles, particularly under Austrian data protection law and the General Data Protection Regulation (GDPR).
- You must obtain **explicit, informed, and unambiguous consent** from individuals before using their data.
- Consent must be specific to the purpose (e.g., machine learning training) and revocable at any time.

[The European Code of Conduct for Research Integrity]
[CARE Principles for Indigenous Data Governance]

Legal framework and good practice

02

How can others use my data?

- sharing according to FAIR principles
- documentation & licensing

- Which rights apply & who is the rights holder?
- How am I allowed to share my data?

➢ **Observe applicable legislation, contracts and policies.**

Licensing

- Provide clear information about the usage modalities of a copyrighted work with
- Can only be granted by the rights holder or someone acting on their behalf
- Different licenses/statements for different types of works
  - Creative Commons for copyrights/related rights of works of art & literature
  - MIT, GNU GPL for software
  - Open Data Commons for databases
  - Europeana Rights Statements for cultural heritage data

- [Choose a license]
- [How to License Research Data (Ball 2014)]
- [How to select an accurate rights statement by Europeana Copyright Community]
- [CLARIN list of license chooser tools]

Other rights

If your model does not fulfil requirements for copyright (intellectual originality and creativity), other protection might still apply e.g. patent rights.

If working at ÖAW you can ask Knowledge Transfer Office for advice.

Open Science

- Open Science paradigms call for greater openness of all steps of the research process, including its outputs.

➤ Choose an open license!

- [FOSTER Open Science]
- [Bezjak et al. 2018]
- [GLAM-E Lab]
- [Ethics and Legality in the Digital Arts and Humanities]

Document

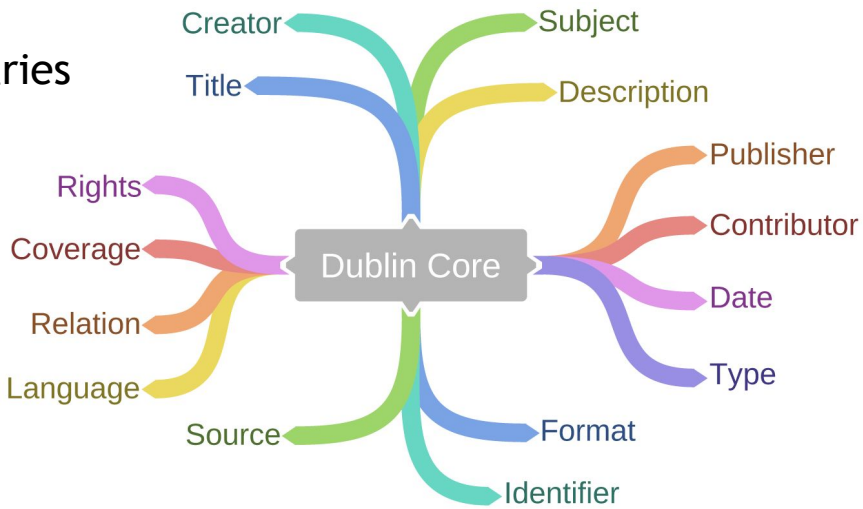
Documentation

- What did you do?
  - How is your data structured?
  - Who did what?
  - What naming conventions are used in your project?
  - How is versioning done?
- README
  - Simple text file (txt)
  - or Markdown (md, CommonMark e.g. GitHub flavour)
    - Example - README.txt IANUS IT-Empfehlungen
    - Example - README.md CORIANDER
    - More on GitHub awesome-readme (with tools listed)

Piotr, CC BY-ND 3.0, via Dataedo Data Cartoons, Kononow 2022

Documentation and Metadata

- A more structured way to describe data, with e.g. ontologies or controlled vocabularies
- Metadata schemas
  - Dublin Core, DataCite, Schema.org
- Ontologies
  - Define classes, properties, and relations
  - E.g., PROV-O, ML Schema
- Controlled vocabularies
  - GND, Getty, TaDiRAH, … (-> more on BARTOC)
  - model your own in SKOS!



Metadata and Machine Readability

- Leveraging the Semantic Web standards, in particular RDF
  - Representing semantics in a machine-readable form
- Information delivered as triples
  - Subject - predicate - object
- Easily linkable with other sources (Linked Open Data)
- (Persistent) identifiers / URIs
  - DOI
  - ORCID
  - …

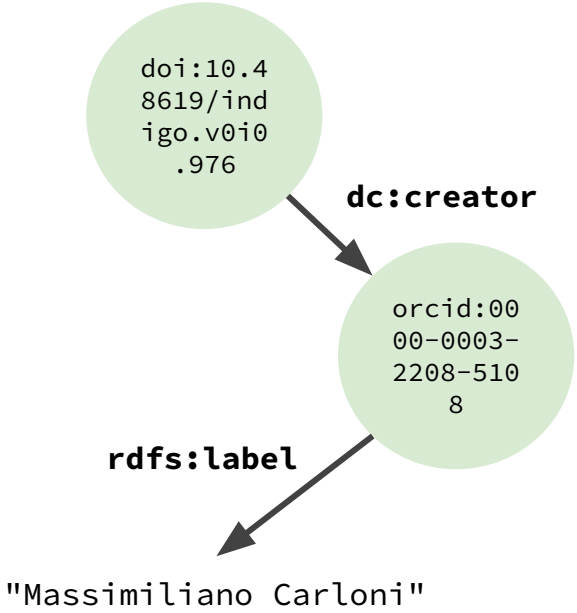
```
<https://doi.org/10.48619/indigo.v0i0.976>
     dc:creator
<https://orcid.org/0000-0003-2208-5108>.

                          |
                          v

<https://orcid.org/0000-0003-2208-5108>
     rdfs:label
"Massimiliano Carloni".
```

Metadata and Machine Readability

- Leveraging the Semantic Web standards, in particular RDF
  - Representing semantics in a machine-readable form
- Information delivered as triples
  - Subject - predicate - object
- Easily linkable with other sources (Linked Open Data)
- (Persistent) identifiers / URIs
  - DOI
  - ORCID
  - …



Metadata and Machine Readability

- Leveraging the Semantic Web standards, in particular RDF
  - Representing semantics in a machine-readable form
- Information delivered as triples
  - Subject - predicate - object
- Easily linkable with other sources (Linked Open Data)
- (Persistent) identifiers / URIs
  - DOI
  - ORCID
  - …

Tim Berners-Lee's
5-star rating system, from
[Linked Data - Design issues]

Creating Metadata for AI Models

- Tools
  - ML Metadata
  - HuggingFace Hub Python library
  - Generators integrated in other tools
  - Auto-generate metadata through custom Python scripts
- Export formats
  - JSON / YAML / …
  - SQL dumps
  - Markdown
- In the next presentation: VELD

```
{
  "model_name": "example-model",
  "version": "1.0",
  "description": "A simple example model for
demonstrating metadata in JSON format.",
  "author": "Your Name",
  "license": "MIT",
  "framework": "PyTorch",
  "task": "Text Classification",
  "inputs": {
    "type": "text",
    "format": "string",
    "max_length": 512
  },
  "hyperparameters": {
    "learning_rate": 0.001,
    "batch_size": 32,
    "epochs": 10
  }, …
```

Model Cards

- Combination of a YAML header followed by Markdown part
  - YAML: more structured and machine-readable metadata
  - Markdown: more explanatory and discursive
- Based on [Mitchell et al. 2019]
- Used in several repositories
  - HuggingFace [docs]
  - Kaggle [docs]
  - GitHub
  - …

```
---
language: en
tags:
- exbert
license: apache-2.0
datasets:
- bookcorpus
- wikipedia
---

# BERT base model (uncased)

Pretrained model on English language using a masked language mod
[this paper](https://arxiv.org/abs/1810.04805) and first release
[this repository](https://github.com/google-research/bert). This
between english and English.

Disclaimer: The team releasing BERT did not write a model card f
the Hugging Face team.

## Model description

BERT is a transformers model pretrained on a large corpus of Eng
was pretrained on the raw texts only, with no humans labeling th
publicly available data) with an automatic process to generate i
was pretrained with two objectives:
```

Model Card for bert-base-uncased

https://huggingface.co/google-bert/bert-base-uncased

https://bit.ly/bert-card

Model Card Template 1

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

from [Mitchell et al. 2019]

Model Card Template 2

Based on [HF Model Card Guidebook]

- Model details
  - Who, when, funding, licence …
- Model sources (opt.)
  - Repo, paper …
- Uses
  - Direct use
  - Downstream use (opt.)
  - Out-of-scope use
- Bias, risks, and limitations
- How to get started with the model
- Training details
  - Training data
  - Training procedure
- Evaluation
  - Testing data, factors & metrics
  - Results
- Model examination (opt.)
- Environmental impact [docs]
- Technical specification (opt.)
- …
- Model card contact
  - Useful to give feedback on the card

Model Card for PowerLM-3B

https://huggingface.co/ibm/PowerLM-3b

https://bit.ly/ibm-power-card

Other Model Cards

https://huggingface.co/Jahid05/llama-3.2-3b-website-prompt-generator

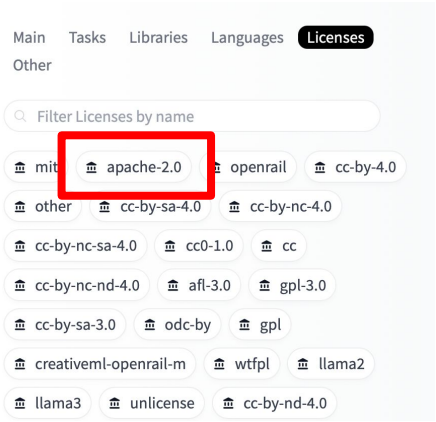
https://bit.ly/wpg-card

https://huggingface.co/arcee-ai/raspberry-3B

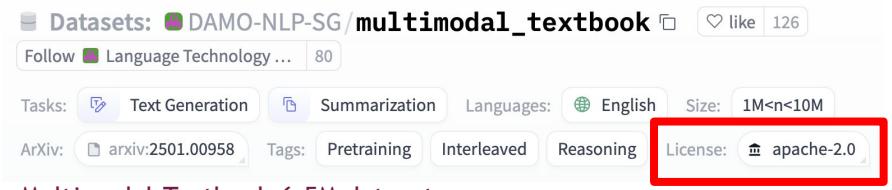
https://bit.ly/rasp-card

Controlled Vocabularies in Hugging Face

e.g. Licences vocabulary from Hugging Face



Dataset search



Multimodal-Textbook-6.5M dataset

Fullname	License identifier (to use in repo card)
Apache license 2.0	`apache-2.0`
MIT	`mit`
OpenRAIL license family	`openrail`
BigScience OpenRAIL-M	`bigscience-openrail-m`
CreativeML OpenRAIL-M	`creativeml-openrail-m`
BigScience BLOOM RAIL 1.0	`bigscience-bloom-rail-1.0`
BigCode Open RAIL-M v1	`bigcode-openrail-m`
Academic Free License v3.0	`afl-3.0`
Artistic license 2.0	`artistic-2.0`
Boost Software License 1.0	`bsl-1.0`

Datasheets for Datasets

Based on [Gebru et al. 2021]

- Motivation
  - Purpose, actors, funding
- Composition
  - Overview of the instances
  - Recommended data splits?
  - Confidential data?
- Collection process
  - Acquisition mechanisms
  - Compensation of workers
  - Ethical review processes
  - Consent from sources
- Preprocessing/cleaning/labeling
  - Were the raw data saved?
  - Is the software used available?
- Uses
  - Already used in some tasks?
  - Impact on future uses
  - Not intended uses
- Distribution
  - Whether/how it will be distributed
  - Licence and/or terms of use
  - Other restrictions
- Maintenance
  - Plans for maintenance/updates
  - Possibility to extend/contribute?

Dataset Cards

Hugging Face version of Datasheets

YAML + Markdown, just like Model Cards

- Dataset details
- Uses
- Dataset structure
- Dataset creation
  - Curation rationale, source data, annotation
- Bias, risks, and limitations
- …
- Dataset card contact

- Docs
  - [Create a dataset card]
  - [Dataset card creation guide]
  - [Dataset card template]
  - [Full list of YAML tags]

Defining splits through YAML metadata

```
my_dataset_repository/
├── README.md
├── data.csv
└── holdout.csv
```

```yaml
---
configs:
- config_name: default
  data_files:
  - split: train
    path: "data.csv"
  - split: test
    path: "holdout.csv"
---
```

Dataset Card for SNLI

Stanford Natural Language Inference corpus

https://huggingface.co/datasets/stanfordnlp/snli

https://bit.ly/snli-card

Why Documentation?

- Improve findability & visibility
  - Especially through structured metadata
- Improve reusability
  - In what context was this model/dataset created
  - Is this suitable for my own purposes?
- Keep a personal/public record of your activity
  - You'll thank yourself in three months' time

Preserve & disseminate

Where do you find your data?

bit.ly/data_survey_02

Repositories for Datasets and Models

Hugging Face https://huggingface.co

https://paperswithcode.com/datasets

kaggle https://www.kaggle.com

Repositories for Datasets and Models

https://openml.org

https://archive.ics.uci.edu

UC Irvine
Machine Learning
Repository

Registry of Open Data on AWS

Repositories for Datasets and Models

Google Dataset Search

https://datasetsearch.research.google.com

https://dataportals.org

DataPortals.org

Wikipedia - List of datasets for machine-learning research

DIP

Data Is Plural Newsletter

How to Find Trusted Repositories

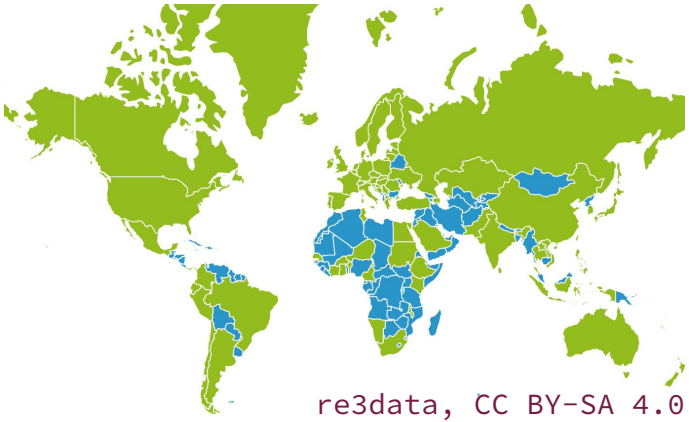
- re3data.org
- OpenDOAR (Open Directory of Open Access Repositories)
- ROAR (Registry of Open Access Repositories)
- FAIRsharing

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

[The TRUST Principles for digital repositories (Lin et al., 2020)]

[CoreTrustSeal]

CORE TRUST SEAL



Examples

Data Archiving and Networked Services
DANS

DANS in Den Haag (NL)

Swedish National Data Service (SWE)

GAMS in Graz (AT)

Examples

https://zenodo.org

- Open and with almost no requirements
- Assigns DOIs
- Communities

https://explore.openaire.eu

- “Parent” initiative of Zenodo
- Discovery platform that aggregates from 141k data sources (incl. Zenodo)

ARCHE

A Resource Centre for the HumanitiEs

arche.acdh.oeaw.ac.at

- Long-term preservation of humanities research data since 2017
- Certified: Core Trust Seal, Clarin B-Centre
- Curation of data and metadata (no self-upload)
- Dissemination
- FAIR principles

CORE TRUST SEAL

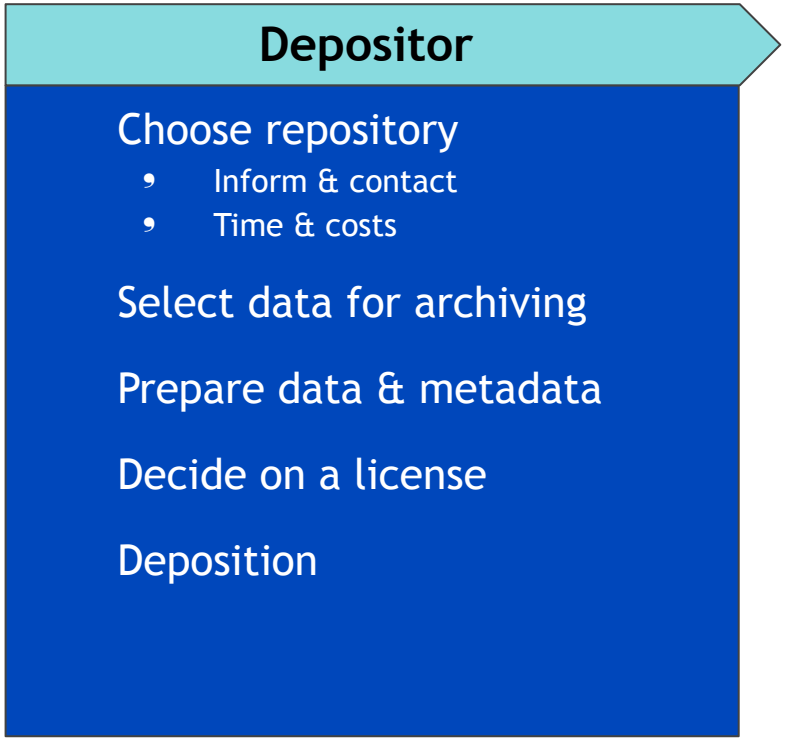
What is ARCHE?

CLARIN B-centre

What does long-term preservation mean?

- > 10 years
- Preservation approaches
  - By migration of formats
  - By emulation of software environments
- ARCHE preservation policy

How do data come into digital archive?



How do data come into digital archive?

Depositor

Choose repository
- Inform & contact
- Time & costs

Select data for archiving

Prepare data & metadata

Decide on a license

Deposit

more data ≠ quality data

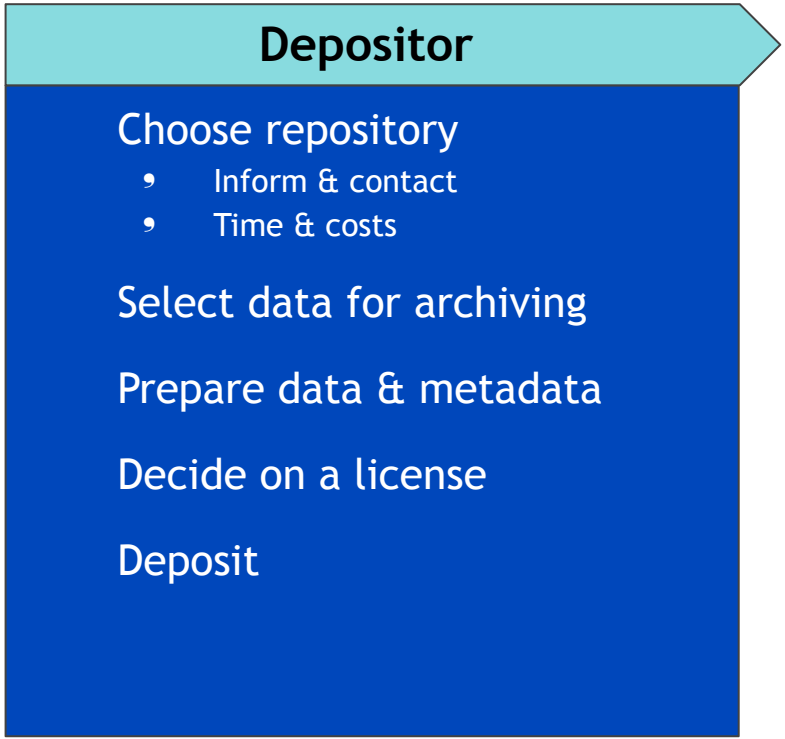
Overabundance of data can cause:

➔ lower discoverability → less reuse

➔ higher costs and duration for curation & archiving

➔ environmental impact of data storage

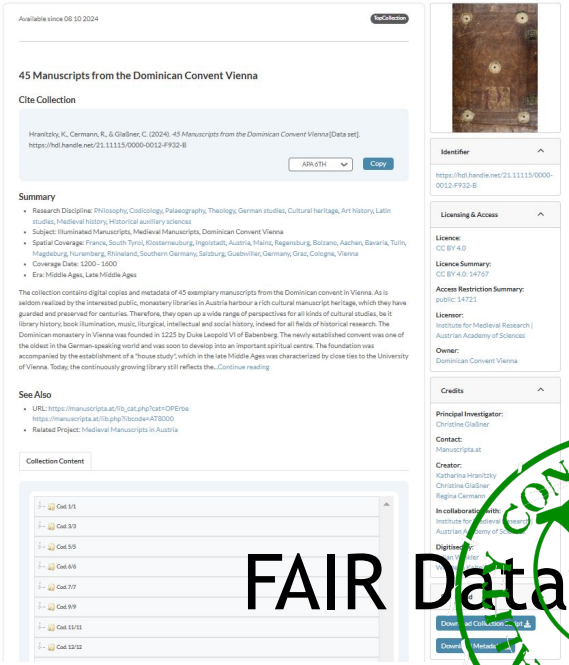
How do data come into digital archive?



How do data come into digital archive?

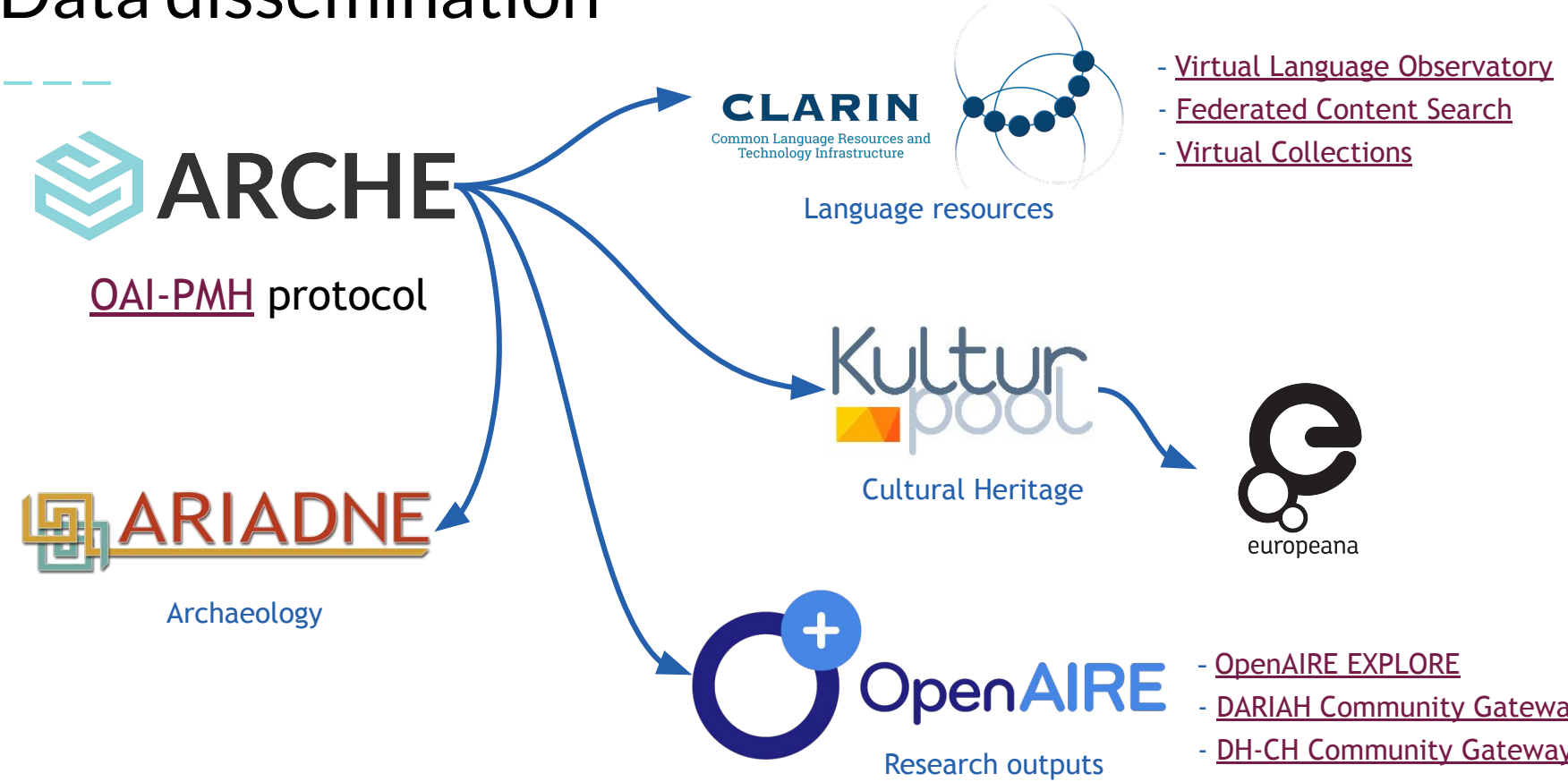
Depositor	Digital Archive
Choose repository • Inform & contact • Time & costs	Deposition Agreement
Select data for archiving	Preform checks • Scope • File validation • Documentation
Prepare data & metadata	Metadata curation & enrichment • ARCHE schema
Decide on a license	
Deposit	Ingestion & dissemination

NOW WHAT?



FAIR Data

Data dissemination



ARCHE APIs

- Programmatic access to datasets and/or metadata
- ARCHE main API with core functionality (CRUD)
- Dissemination Services
  (e.g. to get only EXIF metadata from images)
- You can use ARCHE as back-end for your applications

https://id.acdh.oeaw.ac.at/schnitzler/bahrschnitzler
?format=metadata

OR

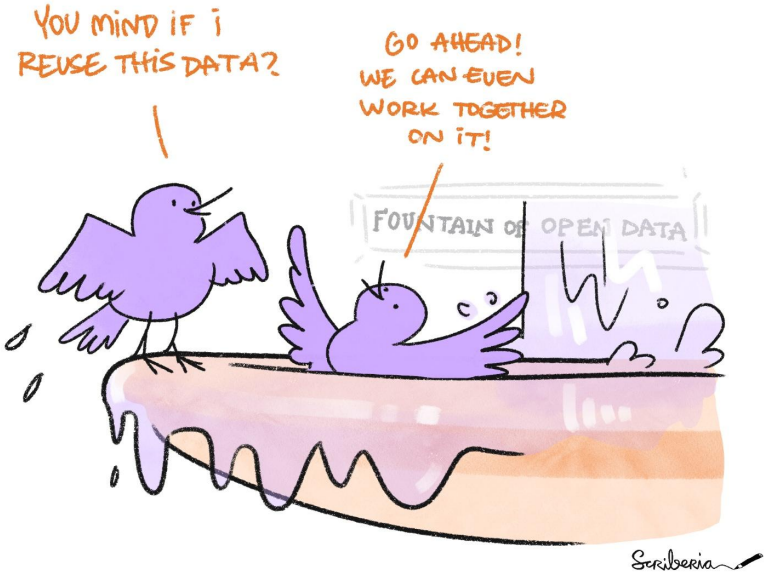
https://hdl.handle.net/21.11115/0000-000E-C8A6-5
@format=metadata

https://arche.acdh.oeaw.ac.at/api/178291

```
Die Korrespondenz Hermann Bahr &ndash; Arthur Schnitzler
        ontology:hasTitle
                "Die Korrespondenz Hermann Bahr – Arthur Schnitzler"@de .
        ontology:hasIdentifier
                id:cmdi/178291 ,
                Die Korrespondenz Hermann Bahr &ndash; Arthur Schnitzler ,
                id:schnitzler/bahrschnitzler ,
                <https://hdl.handle.net/21.11115/0000-000E-C8A6-5> .
        rdfs:type
                ontology:TopCollection .
        ontology:aclRead
                "pandorfer" .
        ontology:aclWrite
                "pandorfer" .
        ontology:createdBy
                "pandorfer" .
        ontology:hasAccessRestrictionSummary
                "öffentlich: 1363"@de^^http://www.w3.org/1999/02/22-rdf-syntax-ns#langString ,
                "public: 1363"@en^^http://www.w3.org/1999/02/22-rdf-syntax-ns#langString .
        ontology:hasActor
                Arthur Schnitzler ,
                Bahr, Hermann .
        ontology:hasAvailableDate
                "2022-04-19T18:02:13.592212"^^http://www.w3.org/2001/XMLSchema#dateTime .
        ontology:hasBinarySize
                "187849959"^^http://www.w3.org/2001/XMLSchema#decimal .
        ontology:hasContact
                Martin Anton M&uuml;ller ,
                Austrian Centre for Digital Humanities and Cultural Heritage .
```

Key takeaways

- Plan data management in advance and revise during the project.
- Keep the data lifecycle going!
- Adhere to FAIR data principles [FAIR for AI].
- Share and reuse, don’t let data go to waste.



Birds of Open Data. *The Turing Way* project illustration by Scriberia. CC-BY 4.0 DOI: 10.5281/zenodo.3332807.

Questions?

Massimiliano Carloni
massimiliano.carloni@oeaw.ac.at

Seta Štuhec
seta.stuhec@oeaw.ac.at

Martina Trognitz
martina.trognitz@oeaw.ac.at

Source and, if available, licence of external illustrations are indicated.

Sources & More

Amnesia

- Anonymization Tool by OpenAIRE [16.01.2025]

ARCHE

- ARCHE & ARCHE Formats. [31.10.2021]

Cambridge Dictionary

- Data [31.10.2021]

CESSDA Training

- CESSDA, Data Management Expert Guide. DOI: 10.5281/zenodo.3820473 [31.10.2021]

Sources & More

DCC

- Digital Curation Coalition (DCC), Data Management Plans. [31.10.2021]

DFG Leitlinien

- DFG, Umgang mit Forschungsdaten. [31.10.2021] (Older version from Jan. 2021: Internet Archive)

DMPonline

- DMPonline, DMP Templates. [31.10.2021]

FAIR Data Principles

- FORCE11, The FAIR Data Principles. [31.10.2021]

Forschungsdaten.info

- forschungsdaten.info [31.10.2021]
  - Der Datenmanagementplan
  - Forschungsdaten veröffentlichen? (eng. publish research data?): A decision tree to guide through important legal aspects when publishing research data. Available in German only.

Sources & More

FOSTER Open Science

- FOSTER, What is Open Science? [15.01.2025]

Fuchs 2019

- S. Fuchs, How do I use 5S method for organizing data files?. In: Think Open Blog, Helsinki. 2019. [15.01.2025]

FWF

- FWF, Forschungsdatenmanagement. [15.01.2025]

GO FAIR

- GO FAIR. [15.01.2025]

Sources & More

IANUS IT-Empfehlungen

- IANUS - M. Heinrich - F. Schäfer - M. Trognitz (Hrsg.), IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften DOI: 10.13149/000.111000-a
  - **IANUS Dateiformate**
    F. Schäfer, Dateiformate.
  - **IANUS Dateiverwaltung**
    M. Trognitz - F. Schäfer - R. Göldner - T. Schenk, Dateiverwaltung.
  - **IANUS Datenmanagement**
    M. Trognitz - S. Jahn - D. Hagmann - J. Räther, Datenmanagement.
  - **IANUS Lebenszyklus**
    M. Trognitz, Der Lebenszyklus von Forschungsdaten.
  - **IANUS Versionskontrolle**
    M. Trognitz, Versionskontrolle

Jones & Grootveld 2017

- S. Jones - M. Grootveld, How FAIR are your data? 2017. [31.10.2021]

Sources & More

Müller-Birn

- C. Müller-Birn, Softwarenutzungsmuster in der Digitalen Forschungsarbeit. 2016. [31.10.2021]

ÖFOS 2012

- Statistik Austria, ÖFOS 2012. Österreichische Version der 'Fields of Science and Technology (FOS) Classification'. 2012. https://vocabs.acdh.oeaw.ac.at/oefosdisciplines/Schema [31.10.2021]

OpenAIRE & EUDAT

- OpenAIRE - EUDAT, ARGOS. [31.10.2021]

Schöch 2013

- C. Schöch, Big? Smart? Clean? Messy? Data in the Humanities. Journal of Digital Humanities 2/3. 2013 [31.10.2021]

Schumm & Steeb 2021

- I. Schumm - L. Steeb, Research Data Management in a Nutshell. BERD@BW Data Literacy Snack, 26.5.2021 [31.10.2021]

Sources & More

Science Europe

- Science Europe, Practical Guide to the International Alignment of Research Data Management. 2019. [31.10.2021]

semver

- T. Preston-Werner, Semantic Versioning 2.0.0. [31.10.2021]

semverdoc

- N. Tekampe, Semantic Versioning for Documents and Meaningful Manual Version Control. [31.10.2021]
  - **semver - MMVC**
    N. Tekampe, Meaningful Manual Version Control Version 1.1.0. [31.10.2021]

Stanford Libraries

- Stanford Libraries, Name files. [31.10.2021]
- Stanford Libraries, Version files. [31.10.2021]

Sources & More

UK Data Service

- UK Data Service, The importance of managing and sharing data (includes research data lifecycle). [31.10.2021]

Wuttke 2019

- U. Wuttke, “Here be dragons”: Open Access to Research Data in the Humanities. 2019. [31.10.2021]